Pergamon

Appl. Math. Lett. Vol. 8, No. 6, pp. 83–87, 1995

0893-9659/95 $9.50 + 0.00

0893-9659(95)00090-9

Integrating Noisy Data

T. PRVAN
Department of Statistics, University of New South Wales
Sydney 2052, Australia

(Received and accepted March 1995)

Abstract—Suppose the parametric form of a curve is not known, but only a set of observations. Quadrature formulae can be used to integrate a function only known from a set of data points. However, the results will be unreliable if the data contains measurement errors (noise). The method presented here fits an even degree piecewise polynomial to the data where all the data points are being used as knot points and the smoothing parameter is optimal for the indefinite integral of the curve which happens to be a smoothing spline. After the smoothing parameter has been chosen, this approach is less computationally expensive than fitting a smoothing spline and integrating.

Keywords—Smoothing spline, Kalman filter, Fixed-interval, Discrete-time smoother, Interpolation smoother.

1. INTRODUCTION

If you did not know the function to integrate, but have values of it specified at four or more data points, say $(t_1, \frac{dy_1}{dt}), \ldots, (t_n, \frac{dy_n}{dt})$, where $\frac{dy_i}{dt}$ denotes $\frac{dy}{dt}$ evaluated at t_i, then quadrature formulae like that in [1] can be used to evaluate the definite integral $\int_{t_1}^{t_n} \frac{dy(x)}{dx}\, dx$. The implicit assumption made is that the data $\frac{dy_1}{dt}, \ldots, \frac{dy_n}{dt}$ at the data points $t_1, \ldots, t_n$ contain no errors. This is not usually the case when data is observed. It is not possible to evaluate $\int_{t_1}^{t} \frac{dy(x)}{dx}\, dx$ at a point t not equaling a data point using quadrature formulae, but in the case of [1] only to $t = t_i,\ i = 4, \ldots, n$ and the information at the data points t_{i+1} to t_n is not utilised.

One approach to integrating data that is noisy is to fit a smoothing spline to the data and then integrate it using the initial conditions that the integrals from t_i to $t_i,\ i = 1, \ldots, n$ are zero. The smoothing parameter is optimal for the smoothing spline fitted and not its indefinite integral, so over smoothing will occur. This approach does permit the calculation of $I = \int_{t_1}^{t} f'(x)\, dx$ at points t not equal to data points and does utilise all the data unlike the quadrature approach. Implicit is the assumption that the data can be decomposed as $\frac{dy_i}{dt} = f'(t_i) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, where the ϵ_i's are independent.

The approach presented here has the indefinite integral being a smoothing spline and the smoothing parameter is optimal for the indefinite integral. This approach also permits the evaluations of definite integrals from t_1 to t where t does not need to coincide with a data point. The amount of work to get from the piecewise curve of degree $2p-2$ with $2p-3$ continuous derivatives fitted to the data to the definite integral $I = \int_{t_1}^{t} f'(x)\, dx$ is negligible, unlike the approach above.

2. INTEGRATING NOISY DATA

Assume that the data $(t_1, y_1), \ldots, (t_1, y_n)$ are observed and not $(t_1, \frac{dy_1}{dt}), \ldots, (t_n, \frac{dy_n}{dt})$ for the time being. A smoothing spline is the solution to the problem of minimizing the following

AML 8-6-G

functional:

$$\sum_{i=1}^{n}(y_i - f(t_i))^2 + \mu \int_{t_1}^{t_n} \left(f^{(p)}(t)\right)^2 dt \tag{1}$$

over f where the data is assumed to be decomposed as

$$y_i = f(t_i) + \epsilon_i, \qquad \epsilon_i \sim N\left(0, \sigma^2\right), \quad i = 1, \ldots, n.$$

The resultant curve is a piecewise polynomial of degree $2p-1$ with $2p-2$ continuous derivatives.

Wecker and Ansley [2] presented a stochastic formulation of the smoothing spline utilising a result by Wahba [3]. She showed that the polynomial smoothing spline is the solution to the stochastic differential equation

$$\frac{d^p x}{dt^p} = \sigma\sqrt{\lambda}\frac{d\omega}{dt},$$

where $\omega(t)$ is a Wiener process (see, for example, [4]) with unit dispersion parameter, $\lambda = 1/\mu$ and $\boldsymbol{x}(t_1) = [x(t_1), \ldots, x^{(p-1)}(t_1)]^\top$ has a diffuse prior (i.e., $\boldsymbol{x}(t_1) \sim N(\mathbf{0}, \gamma^2 I)$ and $\gamma^2 \to \infty$). The solution is

$$x(t) = \lim_{\gamma^2 \to \infty} x(t \mid n).$$

The quantity $x(t \mid n)$ is the expected value of $x(t)$ conditioned on the data $y_1, \ldots, y_n$. The stochastic differential equation can be written in the matrix companion form

$$\frac{d\boldsymbol{x}(t)}{dt} = \begin{pmatrix} \mathbf{0}_{p-1} & I_{p-1} \\ 0 & \mathbf{0}_{p-1}^\top \end{pmatrix} \boldsymbol{x}(t) + \sigma\sqrt{\lambda}\begin{pmatrix} \mathbf{0}_{p-1} \\ \frac{d\omega}{dt} \end{pmatrix}, \tag{2}$$

where $\boldsymbol{x}(t) \in \Re^p$. The fundamental matrix solution of the associated homogeneous differential equation denoted by $T(t_i, t_1)$ has its $(j,k)^{\text{th}}$ element given by

$$T(t_i, t_1)_{jk} = \begin{cases} 0, & j > k, \\ \dfrac{(t_i - t_1)^{k-j}}{(k-j)!}, & j \le k. \end{cases}$$

The observations can be formulated as a signal plus noise model

$$y_i = \boldsymbol{e}_1^\top \boldsymbol{x}_i + \epsilon_i, \qquad \epsilon_i \sim N\left(0, \sigma^2\right), \tag{3}$$

where $\boldsymbol{x}_i = \boldsymbol{x}(t_i)$, the notation $\boldsymbol{e}_j$ is used to denote a p-vector having all zeros except for a one in the j^{th} position and equation (2) can be written recursively as

$$\boldsymbol{x}_i = T_i \boldsymbol{x}_{i-1} + \boldsymbol{u}_i, \tag{4}$$

for $i = 2, \ldots, n$ where T_i is the abbreviation for $T(t_i, t_{i-1})$, $\boldsymbol{u}_i$ is the abbreviation for $\boldsymbol{u}(t_i, t_{i-1})$ which is normally distributed with zero mean, and covariance $\sigma^2\lambda\Omega(t_i, t_{i-1})$ where

$$\Omega(t_i, t_{i-1}) = \int_{t_{i-1}}^{t_i} T(t_i, s)\boldsymbol{e}_p\boldsymbol{e}_p^\top T(t_i, s)^\top ds. \tag{5}$$

Note that Ω_i will be now be used as the abbreviation for $\Omega(t_i, t_{i-1})$.

The dimension of all vector quantities is p unless denoted otherwise, and the elements are real. These equations can be written in matrix form and a likelihood can be associated with the noise vector. This likelihood can be maximized to find the optimal smoothing parameter λ after conditioning on σ^2. Wecker and Ansley [2] avoided working explicitly with the diffuse prior by

obtaining an estimate of the initial state vector in a least squares context and setting the covariance of the initial state vector to be the zero matrix (this corresponds to having no information). The Kalman filter is used as a computational tool in the search for the smoothing parameter. The Kalman filter, fixed-interval, discrete-time smoother, and interpolation smoother are implemented on the state space formulation (3) and (4) to obtain $\boldsymbol{x}(t \mid n)$, and hence, the smoothing spline and its first $p-1$ derivatives evaluated at t; that is, $\boldsymbol{x}(t \mid n) = \left[f(t), f'(t), \ldots, f^{(p-1)}(t)\right]^\top$. Numerically stable and efficient algorithms for these recursions can be found in [5–8].

Osborne and Prvan [6] generalised Wecker and Ansley's [2] stochastic setup to produce a family of curves that included smoothing splines as a special case. The setup used was

$$y_i = \boldsymbol{h}^\top \boldsymbol{x}_i + \epsilon_i, \tag{6}$$

$$\boldsymbol{x}_i = T_i \boldsymbol{x}_{i-1} + \boldsymbol{u}_i, \tag{7}$$

where ϵ_i is normally distributed with zero mean and variance σ^2, and $\boldsymbol{u}_i$ is normally distributed with zero mean and covariance $\sigma^2 \lambda \Omega_i$. The vector quantities belong to $\Re^p$ and the matrix Ω_i is given by

$$\Omega_i = \int_{t_{i-1}}^{t_i} T(t_i, s) V T(t_i, s)^\top ds,$$

where $V : \Re^p \to \Re^p$ is semipositive definite. The relevant recursions to obtain λ and $\boldsymbol{x}(t \mid n)$ are performed now, using (6) and (7) as the the state space formulation instead of (3) and (4). If a diffuse prior is used explicitly (i.e., set $\boldsymbol{x}(t_1) = \boldsymbol{0}$ and $S_{1|0} = \gamma^2 I_p$; note that I_p here denotes the $p \times p$ identity matrix, where γ is chosen to be some large number like 10,000, which from practice works), then the curve fitted is

$$f(t) = \begin{cases} \boldsymbol{h}^\top T(t, t_1) \boldsymbol{x}(t_1 \mid n), & t < t_1, \\ \boldsymbol{h}^\top \boldsymbol{x}(t \mid n), & t_1 \le t \le t_n, \\ \boldsymbol{h}^\top T(t, t_n) \boldsymbol{x}(t_n \mid n), & t > t_n. \end{cases}$$

Here, generalized cross validation or maximum likelihood estimation of a modified likelihood can be used to obtain the smoothing parameter. An efficient algorithm for generalized cross validation can be found in [9]. Note that $\boldsymbol{x}(t_1 \mid n) = \boldsymbol{0}$. When the diffuse prior is dealt with implicitly, the curve fitted is

$$f(t) = \begin{cases} \boldsymbol{h}^\top T(t, t_1) \hat{\boldsymbol{\alpha}}, & t < t_1, \\ \boldsymbol{h}^\top T(t, t_1), \hat{\boldsymbol{\alpha}} + \boldsymbol{h}^\top \boldsymbol{z}(t \mid n), & t_1 \le t \le t_n, \\ \boldsymbol{h}^\top T(t, t_1) \hat{\boldsymbol{\alpha}} + \boldsymbol{h}^\top T(t, t_n) \boldsymbol{z}(t_n \mid n), & t > t_n, \end{cases}$$

where $\hat{\boldsymbol{\alpha}}$ is the estimate of the initial state vector. The estimate of the initial state vector along with its covariance being set to the zero matrix is used to initiate the Kalman Filter. The notation $\boldsymbol{z}(t \mid n)$ is used instead of $\boldsymbol{x}(t \mid n)$ to indicate that different initial conditions have been used. Refer to [2] for more details. In [6], the quantities V and $\boldsymbol{h}$ were shown to have a role to play in the resultant continuity properties of the curve fitted. What was not discussed is that $\boldsymbol{h}$ also has a role to play in determining the degree of the polynomial being fitted with the maximum degree possible being $2p-1$.

The derivative of a polynomial smoothing spline of degree $2p-1$ is the second element of the state vector in the stochastic formulation of the smoothing spline; that is, $\boldsymbol{e}_2^\top \boldsymbol{x}(t \mid n)$. Now consider the data $(t_1, \frac{dy_1}{dt}), \ldots, (t_n, \frac{dy_n}{dt})$, at the data point t_i the observation equation can be expressed as

$$\frac{dy_i}{dt} = \boldsymbol{e}_2^\top \boldsymbol{x}_i + \epsilon_i', \qquad \epsilon_i' \sim N\left(0, \sigma'^2\right), \tag{8}$$

where $\boldsymbol{x}_i$ still obeys the relationship

$$\boldsymbol{x}_i = T_i \boldsymbol{x}_{i-1} + \boldsymbol{u}_i, \qquad \boldsymbol{u}_i \sim N(\boldsymbol{0}, \sigma^2 \lambda \Omega_i), \tag{9}$$

for $i = 2, \ldots, n$ and Ω_i is given by equation (5). If the covariance of $\boldsymbol{u}_i$ was now thought of as $\sigma'^2\lambda'\Omega_i$, then we could proceed as Osborne and Prvan [6] to produce a polynomial of degree $2p-2$ with $2p-3$ continuous derivatives. The curve being fitted is $\boldsymbol{e}_2^\top \boldsymbol{x}(t \mid n)$ when the diffuse prior is used explicitly and $\frac{dy_i}{dt}, i = 1, \ldots, n$ are the data points. The first element of the smoothed state vector $\boldsymbol{x}(t \mid n)$ will give an integral of the curve being fitted from t_1 to t. We can get an idea of what the family of curves that solves the differential equation $\frac{dy}{dt} = f'(t)$ looks like. The smoothing parameter is optimal for the integral and not for the curve fitted to the data.

If the Wecker and Ansley [2] approach is used to obtain the optimal smoothing parameter, then when solving for the initial state vector, the associated matrix will be rank deficient by 1. The system of equations to solve is of the form

$$\mathrm{L}^{-1}\begin{pmatrix} \boldsymbol{e}_2^\top \\ \boldsymbol{e}_2^\top T(t_2, t_1) \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{e}_2^\top T(t_n, t_1) \end{pmatrix} \boldsymbol{\alpha} = L^{-1}\begin{pmatrix} \frac{dy_1}{dt} \\ \frac{dy_2}{dt} \\ \cdot \\ \cdot \\ \cdot \\ \frac{dy_n}{dt} \end{pmatrix}.$$

The matrix L is the square root of the covariance matrix associated with the state space formulation (8) and (9) being written in matrix form. The first column of the matrix above contains all zeros. The system of equations can be solved by a singular value decomposition, where $\boldsymbol{\alpha} = \boldsymbol{x}(t_1)$. The calculations can be simplified by applying a permutation which swaps the first column with the last and then doing least squares. When applying Householder matrices to the permuted system of equations, the system can be transformed to upper triangular form with the last column of the influence matrix all zeros, the least squares solution is obtained by setting the p^{th} element to zero (which corresponds to the first element in the original setup) and solving for the other $p-1$ elements. This is called the basic solution, and in this special case, corresponds to the singular value decomposition solution because of the zeroes to the right. The residual sum of squares is just the sum of the squares of the last n-p elements on the right-hand side. The estimate of the initial state vector will have a zero in the first position, which is consistent with the integral of the curve being fitted from t_1 to t_1 being zero. For further details on the singular value decomposition, see [10].

When $\boldsymbol{h} = \boldsymbol{e}_1$ and $V = \boldsymbol{e}_p\boldsymbol{e}_p^\top$ in our state space formulation (6) and (7), we have a smoothing spline. If $\boldsymbol{h} = \boldsymbol{e}_2$ is used instead, we are still minimizing the functional (1), but the curve we are fitting is an estimate of the data, $\frac{dy_i}{dt}$, $i = 1, \ldots, n$, given by $f'(t)$, so the resultant curve is a polynomial of degree $2p-2$ with $2p-3$ continuous derivatives. When we are dealing with the diffuse prior explicitly, the first entry of $\boldsymbol{x}(t \mid n)$ gives the integral of $f'(t)$ from t_1 to t, and this integral is a smoothing spline.

3. SIMULATION STUDY

The function $f'(t) = 3t^2(t^2+1)^{-1} - 2t^4(t^2+1)^{-2}$ was considered, which has indefinite integral $t^3(t^2+1)^{-1} + C$. The data $\frac{dy_i}{dt} = f'(t_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, 25$, were simulated 100 times for each $\sigma = 0.01$, $\sigma = 0.1$ and $\sigma = 1.0$ with the data points -3.8, -3.1, -2.7, -2.5, -1.9, -1.6, -1.1, -0.9, -0.7, -0.4, -0.3, -0.2, -0.1, 0.0, 0.1, 0.2, 0.3, 0.5, 0.8, 1.2, 1.8, 2.3, 2.9, 3.1, 3.5. For each value of σ, the three approaches for obtaining an estimate of $\int_{t_1}^{t_i} f'(t)\,dt$, and an estimate of the variability. The results are given below where method 1 is the Gill and Miller [1] quadrature method, method 2 is integrating a cubic smoothing spline that has been fitted to the data with the smoothing parameter chosen by generalised cross validation, and method 3 is that outlined in Section 2 using generalised cross validation to obtain the smoothing parameter ($p = 2$). The quantity $\int_{t_1}^{t_i} \hat{f}_j'(t)\,dt$ represents the approximation to $\int_{t_1}^{t_i} f'(t)\,dt$ for the j^{th} simulated data set for the σ given.

Table 1. Comparison of the three methods.

σ	i	$\max_j \left\lvert \frac{\int_{t_1}^{t_i} f'(t)\,dt - \int_{t_1}^{t_i} \hat{f}'_j(t)\,dt}{\int_{t_1}^{t_i} f'(t)\,dt} \right\rvert \times 100\%$			$\sum_{j=1}^{100} \frac{\left(\int_{t_1}^{t_i} f'(t)\,dt - \int_{t_1}^{t_i} \hat{f}'_j(t)\,dt\right)^2}{100}$		
		Method 1	Method 2	Method 3	Method 1	Method 2	Method 3
0.01	25	2.96 %	1.56 %	1.62 %	9.7×10^{-3}	1.2×10^{-3}	1.9×10^{-3}
	19	3.46 %	1.70 %	1.47 %	3.7×10^{-3}	7.7×10^{-4}	5.9×10^{-4}
	13	6.42 %	2.71 %	2.03 %	1.2×10^{-2}	1.6×10^{-3}	8.6×10^{-4}
	7	7.44 %	2.66 %	2.03 %	1.1×10^{-2}	1.4×10^{-3}	9.1×10^{-4}
0.10	25	29.5 %	15.9 %	13.7 %	9.6×10^{-1}	1.9×10^{-1}	1.3×10^{-1}
	19	34.4 %	17.1 %	15.4 %	3.7×10^{-1}	8.0×10^{-2}	6.6×10^{-2}
	13	64.2 %	27.1 %	22.4 %	1.2×10^{0}	1.5×10^{-1}	1.0×10^{-1}
	7	74.5 %	26.5 %	19.1 %	1.1×10^{0}	1.3×10^{-1}	8.6×10^{-2}
1.0	25	295 %	160 %	147 %	1.1×10^{1}	1.3×10^{1}	1.1×10^{1}
	19	344 %	179 %	145 %	3.7×10^{1}	8.2×10^{0}	5.8×10^{0}
	13	642 %	260 %	227 %	1.2×10^{2}	1.3×10^{1}	9.7×10^{0}
	7	745 %	230 %	195 %	1.1×10^{2}	1.1×10^{1}	7.1×10^{0}

From Table 1, it can be seen that the quadrature method always performs poorly in comparison with the other two methods. When the noise is close to zero ($\sigma = 0.01$), there isn't much difference between fitting a cubic smoothing spline to the data and then integrating or using the method outlined in Section 2. This is consistent with the tendency to interpolate the data. When the noise is noticeable and the signal discernable ($\sigma = 0.1$), the method outlined in Section 2 is superior. When the noise drowns out the signal ($\sigma = 1.0$), methods 2 and 3 are vastly superior to the quadrature method with not much difference between them. This is consistent with the tendency to over smooth the data.

With methods 2 and 3, the main computational effort is in obtaining the smoothing parameter for the data. For the simulations carried out, this took on average 7.66 seconds for method 2 and 7.69 seconds for method 3. For each simulated data set, the smoothing parameter was searched for in the intervals $(10^k, 10^{k+1})$, $k = -6, \ldots, 5$. Once the smoothing parameter has been found, to obtain an estimate of a definite integral from t_1 to some specified t took on average 0.0068 seconds for method 2 and 0.0046 seconds for method 3.

REFERENCES

1. P.E. Gill and G.F. Miller, An algorithm for the integration of unequally spaced data, *Comput. J.* **15**, 80–83 (1972).
2. W. Wecker and C.F. Ansley, Signal extraction approach to nonlinear regression and spline smoothing, *J. Amer. Statist. Assoc.* **78**, 81–89 (1983).
3. G. Wahba, Improper priors, spline smoothing and the problem of guarding against model errors in regression, *J. R. Statist. Assoc.* **B40**, 364–372 (1978).
4. P. Billingsley, *Probability and Measure*, John Wiley and Sons, New York, (1986).
5. M.R. Osborne and T. Prvan, On algorithms for generalised smoothing splines, *J. Austral. Math. Soc.* **B29**, 43–56 (1988).
6. M.R. Osborne and T. Prvan, Smoothness and conditioning in generalised smoothing spline calculations, *J. Austral. Math. Soc.* **B30**, 319–338 (1988).
7. M.R. Osborne and T. Prvan, What is the covariance analog of the Paige and Saunders information filter?, *SIAM J. Sci. Stat. Comput.* **12**, 1324–1331 (1991).
8. T. Prvan and M.R. Osborne, A square-root fixed-interval, discrete-time smoother, *J. Austral. Math. Soc.* **B30**, 57–68 (1988).
9. C.F. Ansley and R. Kohn, Efficient generalized cross-validation for state space models, *Biometrika* **74**, 139–148 (1987).
10. G.H. Golub and C.F. Van Loan, *Matrix Computations*, Second edition, The Johns Hopkins University Press, Baltimore, MA, (1989).